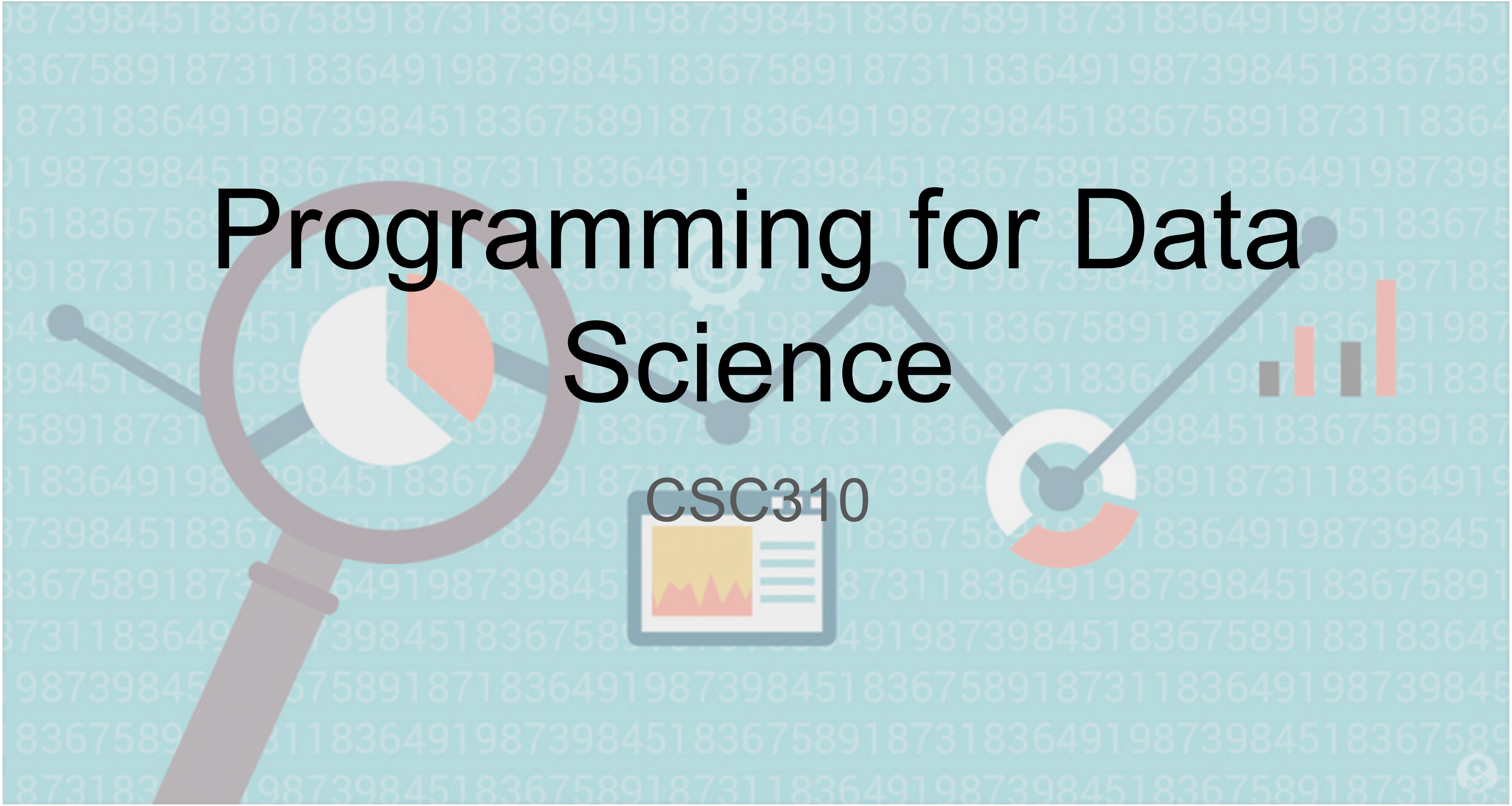
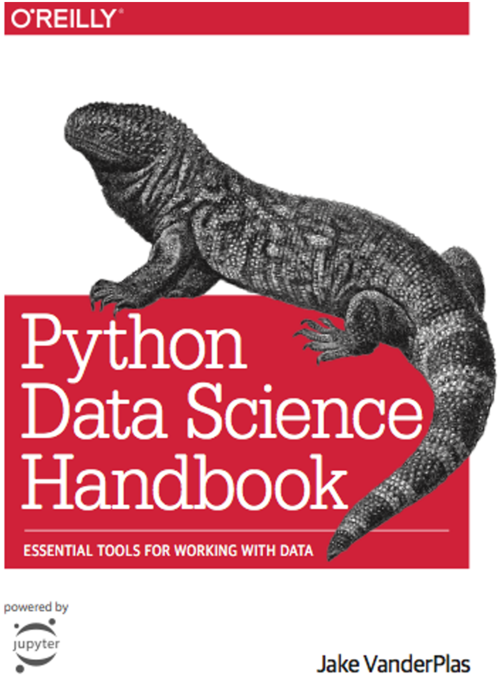
Programming for Data Science

CSC310



Course Details

- Dr Lutz Hamel
- Best way to get in touch - email:
  - lutzhamel@uri.edu
- Everything is online
  - Assignments & Gradebook & Syllabus
    - BrightSpace
  - Lecture Notes
    - https://lutzhamel.github.io/CSC310/
  - Book
    - Python Data Science Handbook

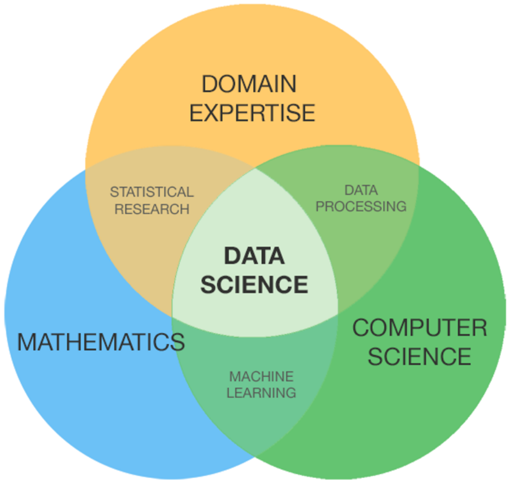


What is Data Science?

☞ **Data science is the discipline of the extraction of knowledge from data.**

It relies on

- computer science
  - for AI, data structures, algorithms, visualization, big data support, and general programming
- statistics/mathematics
  - for data models and inference
- domain expertise
  - for asking questions and interpreting results



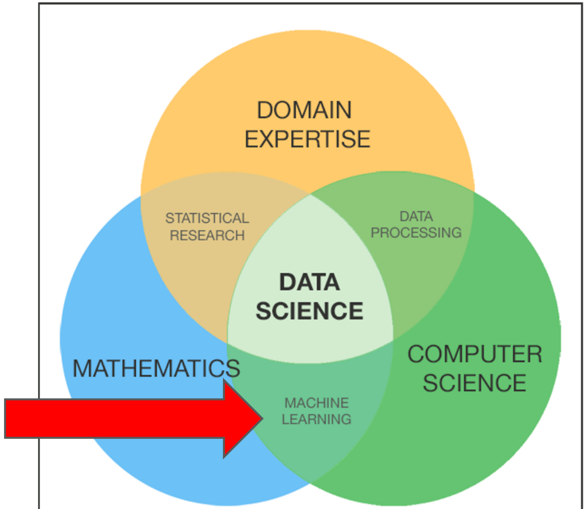
Source: Palmer, Shelly. Data Science for the C-Suite. New York: Digital Living Press, 2015. Print.

What is Data Science?

☞ **Data science is the discipline of the extraction of knowledge from data.**

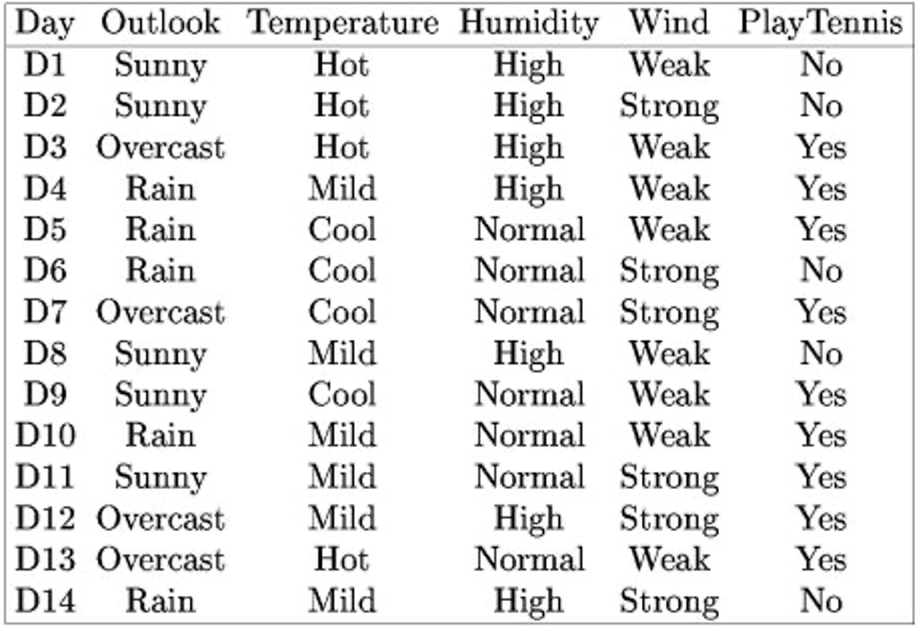
How do we do that?

☞ **We build MODELS of data!**



Source: Palmer, Shelly. Data Science for the C-Suite. New York: Digital Living Press, 2015. Print.

Models: Play Tennis

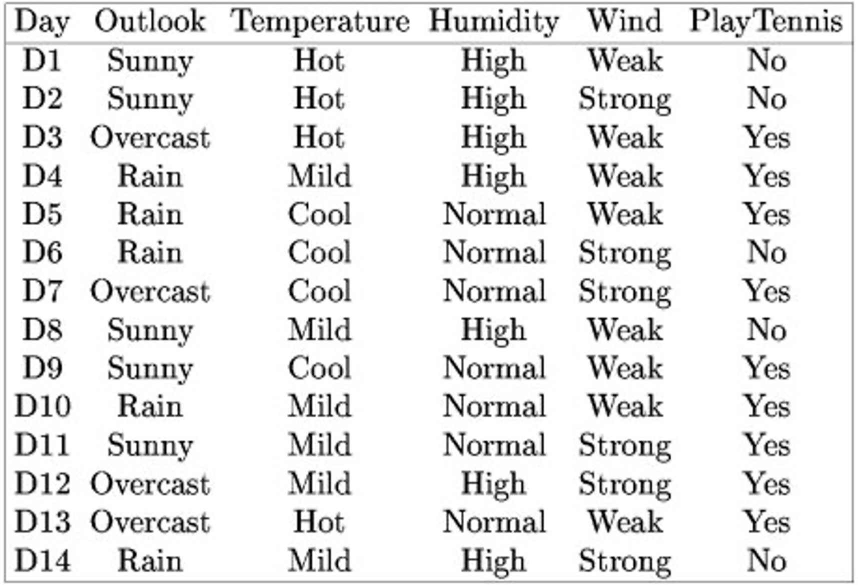


Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Lots of data - very little information!

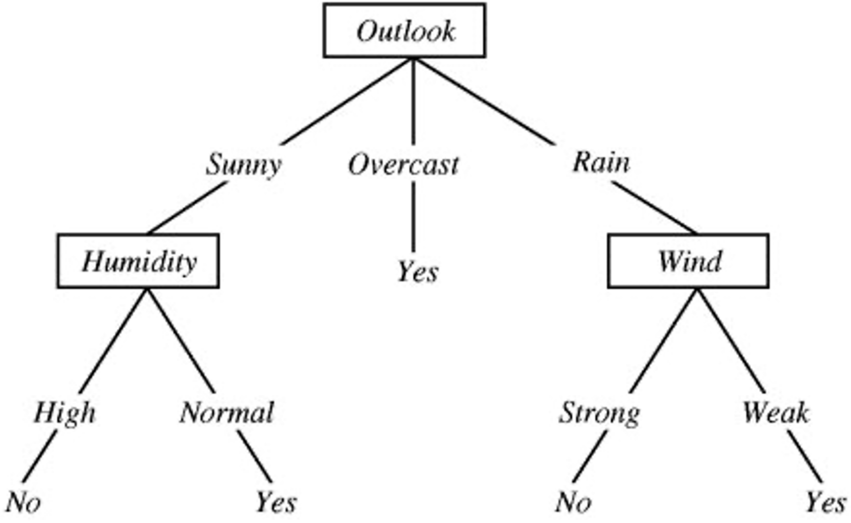
Build a model - **a decision tree!**

Models: Play Tennis



Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

ID3 Decision Tree



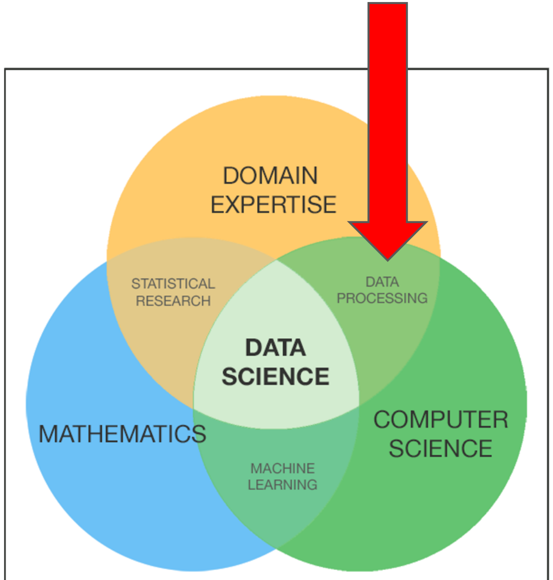
☞This model summarizes the whole table correctly!

What is Data Science?

☞ **Data science is the discipline of the extraction of knowledge from data.**

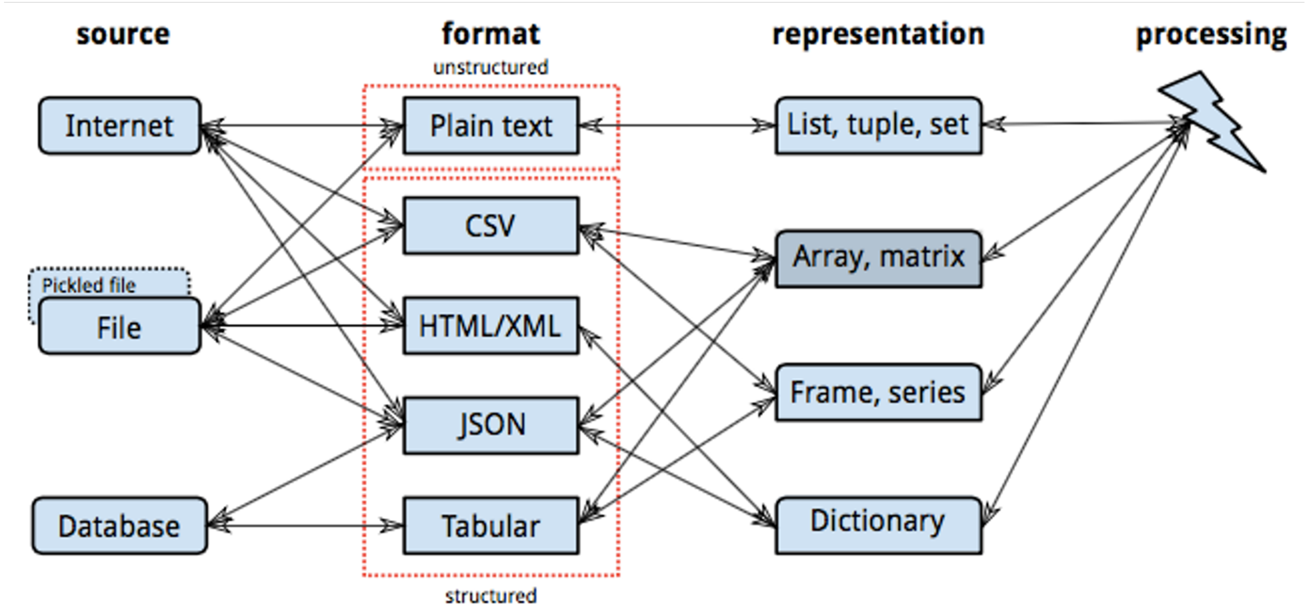
Where does the data come from?

☞ **The data pipeline!**



Source: Palmer, Shelly. Data Science for the C-Suite.
New York: Digital Living Press, 2015. Print.

The Data Pipeline



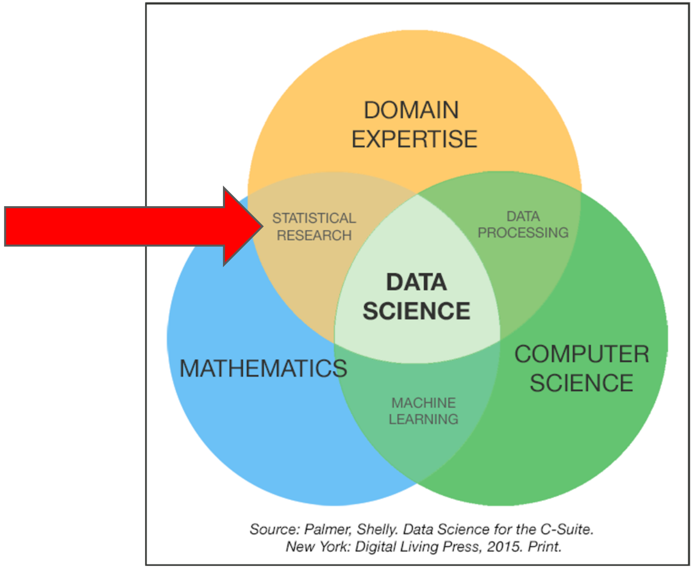
What is Data Science?

☞ **Data science is the discipline of the extraction of knowledge from data.**

How do we preprocess our data for model building?

☞ **Statistics!**

- Descriptive Statistics
- Missing Value Processing
- Normalization



Source: Palmer, Shelly. Data Science for the C-Suite. New York: Digital Living Press, 2015. Print.

Descriptive vs. Inferential Statistics

Purpose: Descriptive statistics aim to summarize data, while inferential statistics aim to make predictions or generalizations about a population from a sample.

Data coverage: Descriptive statistics deal with the entire dataset, whereas inferential statistics focus on samples from which to generalize about a population.

Analysis outcome: The outcome of descriptive statistics is a summary of data, while the outcome of inferential statistics is predictions, decisions, or inferences about population parameters.

In summary, descriptive statistics help describe, show, or summarize data in a meaningful way, allowing the data to be visualized easily, whereas inferential statistics take data from a sample and make inferences or predictions about a population.

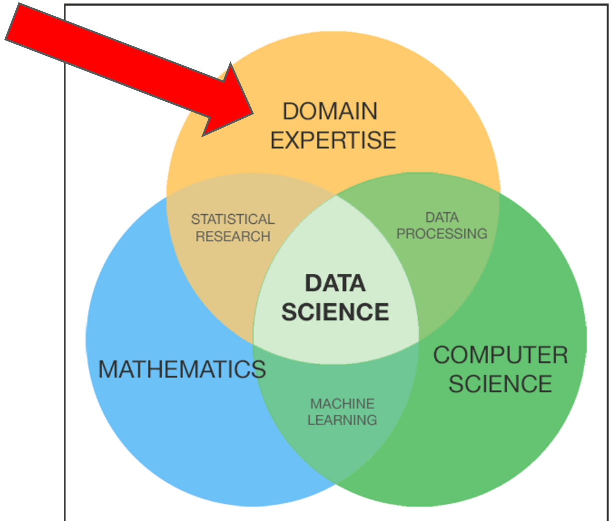
What is Data Science?

☞ **Data science is the discipline of the extraction of knowledge from data.**

How do we ask the right questions?

☞ **Domain Expertise!**

Knowledge cannot be generated in a vacuum. You need the context of a domain in order to generate new insights. E.g. bioinformatics, climate modeling, sales forecasting, *etc.*



Source: Palmer, Shelly. Data Science for the C-Suite. New York: Digital Living Press, 2015. Print.